Research Report

Mixture Distribution Diagnostic Models

Matthias von Davier

Research & Development

July 2007
RR-07-32

Mixture Distribution Diagnostic Models

Matthias von Davier
ETS, Princeton, NJ

July 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper introduces the mixture general diagnostic model (MGDM), an extension of the general diagnostic model (GDM). The MGDM extension allows one to estimate diagnostic models for multiple known populations as well as discrete unknown, or not directly observed mixtures of populations. The GDM is based on developments that integrate located latent class models; multiple classification latent class models; and discrete, multidimensional item response models into a common framework. Models of this type express the probability of a response vector as a function of parameters that describe the individual item response variables in terms of required skills and of indirectly observed (latent) skill profiles of respondents. The skills required for solving the items are, as in most diagnostic models, represented as a design matrix that is often referred to as a Q-matrix. This Q-matrix consists of rows describing, for each item response, what combination of skills is needed to succeed or to obtain partial or full credit. The hypothesized Q-matrix is either the result of experts rating items of an existing assessment (retrofitting) or comes directly out of the design of the assessment instrument, in which it served as a tool to design the items.

The MGDM takes the GDM and integrates it into the framework of discrete mixture distribution models for item response data (see von Davier & Rost, 2006). This increases the utility of the GDM by allowing the estimation and testing of models for multiple populations. The MGDM allows for complex scale linkages that make assessments comparable across populations and makes it possible to test whether items function the same in different subpopulations. This can be done with known subpopulations (defined by grade levels, cohorts, etc.), as well as with unknown subpopulations that need to be identified by the model. In both cases, MGDMs make it possible to determine whether different sets of item-by-skill parameters and/or different skill distributions have to be assumed for different subpopulations. This amounts to a generalized procedure that can be used to test for differential item functioning (DIF) on one item or on multiple-response variables using multiple-group or mixture models. This procedure enables testing DIF models against models that allow additional skills for certain items in order to account for differences between subpopulations.

Key words: Item response theory, diagnostic models, discrete MIRT, mixture distribution models, multiple classification latent class analysis, polytomous data

Acknowledgments

This paper uses examples from joint work with Henry Braun, Alina von Davier, Xiaomin Huang, Xueli Xu, and Kentaro Yamamoto. The opinions presented here are the author's and not necessarily shared by ETS.

What Are Diagnostic Models?

Rule space methodology (Tatsuoka, 1983) and latent structure models with multiple latent classifications (Goodman, 1974a, 1974b; Haberman, 1979; Haertel, 1989; Maris, 1999) represent the most well known early attempts at diagnostic modeling. The noisy-input deterministic-and (NIDA) model (Junker & Sijtsma, 2001; Maris 1999) is an example of a recently discussed diagnostic model. Similarly, the deterministic-input noisy-and (DINA) model, which is a constrained (multiple classification), latent class model has been discussed by several authors (Haertel, 1989; Junker & Sijtsma; Macready & Dayton, 1977). More recently, the unified model (DiBello, Stout, & Roussos, 1995), which lacks identifiability in its original parameterization, underwent modification and was recast as the reparameterized unified model (RUM; also referred to as the *fusion model* or the *Arpeggio system*; Hartz, Roussos, & Stout, 2002).

This paper introduces a class of models for cognitive diagnosis, the general diagnostic model (GDM; von Davier, 2005a), in its form for multiple populations and discrete mixtures. The GDM is based on developments that integrate (located) latent class models (Formann, 1985; Lazarsfeld & Henry, 1968); multiple classification latent class models (Maris, 1999); and discrete, multidimensional item response theory (MIRT) models (Reckase, 1985) into one common framework. von Davier showed that the GDM contains several previous approaches in addition to some common IRT models as special cases. Similar to previous approaches to diagnostic modeling, GDMs describe the probability of a response vector as a function of parameters that describe the individual item response variables in terms of required skills and of indirectly observed (latent) skill profiles of respondents. The item-by-skill requirements are recorded in most diagnostic models as a design matrix that is often referred to as a Q-matrix. This Q-matrix consists of rows representing a hypothesis of the combination of skills needed to succeed or to obtain partial or full credit in response to a particular item. The hypothesized Q-matrix is either the result of experts rating items of an existing assessment (retrofitting) or comes directly out of the design of the assessment instrument, in which it served as a tool to design the items. In this paper, models referred to as the GDM (von Davier, 2005a; von Davier & Yamamoto, 2004a, 2007) will be introduced and extended to mixtures and multiple groups, which this paper refers to under the title of the mixture general diagnostic model (MGDM). GDMs have been developed to integrate multiple-classification, latent class models (Maris, 1999) and located, latent class models (Formann, 1985) and may be described as discrete MIRTs (Lord & Novick, 1968; Reckase, 1985). The MGDM extends the GDM to mixtures of discrete MIRT models (von Davier &

Rost, 2006; von Davier & Yamamoto, 2007). Unidimensional mixture IRT models were described by Mislevy and Verhelst (1990), Kelderman and Macready (1990), and Rost (1990). von Davier and Rost (1995) extended conditional maximum likelihood methods to mixture Rasch models for polytomous data, and von Davier and Yamamoto (2004b) described the mixture distribution generalized partial credit model (mixture GPCM), an extension of the GPCM (Muraki, 1992).

The extension of the GDM to multiple groups and/or mixtures of populations increases the utility of GDMs by making it possible to estimate and test models in such settings. For example, the MGDM allows for complex scale linkages to compare assessments across populations (compare von Davier & von Davier, 2004), and it enables testing whether items are functioning the same in different populations. This can be done with either known populations (grades, cohorts, etc.) or with unknown subpopulations that need to be identified by the model. In both cases, MGDMs make it possible to test whether different sets of item-by-skill parameters and/or different skill distributions have to be assumed for different subpopulations. This amounts to a generalized procedure that can be used to test for DIF on one item or on multiple response variables using multiple-group or mixture models and to test such DIF models against models that identify additional skills for certain items in order to account for differences between subpopulations.

Diagnostic models typically assume a multivariate, but discrete, latent variable that represents the absence or presence, or more gradual levels, of multiple skills. These skill profiles have to be inferred through model assumptions with respect to how the observed data relate to the unobserved skill profile. The absence or presence of skills is commonly represented by a Bernoulli (0/1) random variable in the model. Given that the number of skills represented in the model is larger than in unidimensional models (obviously greater than 2, but smaller than 14 skills in most cases), the latent distribution of skill profiles needs some specification of the relationship between skills in order to avoid the estimation of up to 2^{14}-1 = 16,383 separate skill-pattern probabilities. The GDM (von Davier, 2005a) allows ordinal skill levels and different forms of skill dependencies to be specified so that more gradual differences between examinees can be modeled in this framework.

The following section will introduce the GDM for dichotomous and partial credit data and binary as well as ordinal latent skill profiles. Then the MGDM will be introduced. Third, scale linkage across multiple groups using GDMs will be discussed. Finally, examples of applications of the MGDM in large-scale data analysis will be presented.

The General Diagnostic Model Framework

von Davier and Yamamoto (2004a) developed a GDM framework that uses ideas from MIRT and multiple-classification and located latent class models The GDM is suitable for polytomous items, for dichotomous items, and for mixed items in one or more test forms. It enables the modelling of polytomous skills, mastery/nonmastery skills, and pseudo-continuous skills. von Davier (2005b) described the partial credit GDM and developed an expectation-maximization (EM) algorithm to estimate GDMs. In 2006, this algorithm was extended to the estimation of MGDMs.

General Diagnostic Models for Ordinal Skill Levels

This section introduces the GDM (developed by von Davier & Yamamoto, 2004a) for dichotomous and polytomous data and ordinal skill levels. Diagnostic models can be defined by a discrete, multidimensional, latent variable θ; in the case of the MGDM, the multidimensional skill profile $\vec{a} = (a_1,\ldots a_K)$ consists of discrete, user-defined skill levels $a_k \in \{s_{k1},\ldots, s_{kl},\ldots, s_{kL_k}\}$.

In the simplest (and most common) case, the skills are dichotomous (i.e., the skills will take on only two values $a_k \in \{0,1\}$). In this case, the skill levels are interpreted as mastery (1) versus nonmastery (0) of skill k. Let $\theta = (a_1,...,a_K)$ be a K-dimensional skill profile consisting of K polytomous skill levels $a_k, \quad k = 1,...,K$.

The probability of a response *x* in the general diagnostic model is given by

$$P(X_i = x | \vec{\beta}_i, \vec{q}_i, \vec{\gamma}_i, \vec{a}) = \frac{\exp\left[\beta_{xi} + \sum_{k=1}^{K} \gamma_{xik} h_i(q_{ik}, a_k)\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yi} + \sum_{k=1}^{K} \gamma_{yik} h_i(q_{ik}, a_k)\right]}, \tag{1}$$

with k-dimensional skill profile $a = (a_1,...,a_K)$ and with some necessary restrictions on the $\sum_k \gamma_{xik}$ and the $\sum \beta_{xi}$ to identify the model.

The Q-matrix entries q_{ik} relate item i to skill k and determine whether or not (and to what extent) skill k is required for item i. If skill k is required for item i, then $q_{ik} > 0$. If skill k is not required, then $q_{ik} = 0$.

The real functions $h_i(q_{ik}, a_k)$ are a central building block of the GDM. The function h_i maps the skill levels a_k and Q-matrix entries q_{ik} to the real numbers. In most cases, the

same mapping will be adopted for all items, so one can drop the index i. The h mapping defines how the Q-matrix entries and the skill levels interact (von Davier, 2005a; von Davier, DiBello, & Yamamoto, 2006).

Examples of Skill Level Definitions

Assume that the number of skill levels is $S_k = 2$, and choose skill levels $a_k \in \{-1.0, +1.0\}$, or alternatively $a_k \in \{-0.5, +0.5\}$. Note that these skill levels are a priori defined constants and not model parameters. This setting can be easily generalized to polytomous, ordinal skill levels with the number of levels being $S_k = m+1$ and a determination of levels such as $a_k \in \{(0-c), (1-c), \ldots, (m-c)\}$ for some constant c. An obvious choice is $c = m/2$.

Consider a case with just one dimension, say $K = 1$, and many levels, say $S_k = 41$, with levels of a_k being equally spaced (a common, but not a necessary, choice), say $a_k \in \{-4.0, \ldots, +4.0\}$. Here, the GDM mimics a unidimensional IRT model, namely the GPCM (Muraki, 1992). As a consequence, this IRT-like version of the GDM requires constraints to remove the indeterminacy of the scale, just as IRT models do.

For GDMs with just a few levels per skill, such constraints may not be needed. In the (most) common case of two levels per skill, the range of skill levels is counterbalanced by the average of slope parameters. For example, a GDM with $a_k \in \{-1.0, +1.0\}$ produces slope parameters that are half as big as a GDM that uses $a_k \in \{-0.5, +0.5\}$ as skill levels. This case does not require constraints, as just one proportion determines the mean and variance of a binary variable.

von Davier and Yamamoto (2004a) showed that the GDM already contains a compensatory version of the fusion model as well as many common IRT models as special cases. The parameters β_{xi} may be viewed as item difficulties in the dichotomous case and as threshold parameters in the polytomous case, and the γ_{xik} may be interpreted as slope parameters.

General Diagnostic Models for Partial Credit Data

For a partial credit version of the GDM, choose $h_i(q_{ik}, a_k) = q_{ik} a_k$ with a binary (0/1) Q-matrix. The resulting model contains many standard IRT models and their extensions to confirmatory MIRT models using Q-matrices. This GDM may be viewed as a multivariate,

discrete version of the GPCM. For a response $x \in \{0,1,2,...,m_i\}$, the model based probability in this GDM is

$$P(X_i = x | \vec{\beta}_i, \vec{a}, \vec{q}_i, \vec{\gamma}_i) = \frac{\exp\left[\beta_{xi} + \sum_{k=1}^{K} x\gamma_{ik} q_{ik} a_k\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yi} + \sum_{k=1}^{K} y\gamma_{ik} q_{ik} a_k\right]}, \quad (2)$$

with k attributes (discrete latent traits) $a = (a_1,...,a_K)$ and a dichotomous design Q-matrix $(q_{ik})_{i=1..I,k=1..K}$. These a_k are discrete scores determined before estimation and can be chosen by the user. These scores are used to assign real numbers to the skill levels; for example, $a(0) = -1.0$ and $a(1) = +1.0$ may be chosen for dichotomous skills.

For a vector of item responses, local stochastic independence (LI) is assumed, which yields

$$P(\vec{X} = \vec{x} | \vec{\vec{\beta}}, \vec{a}, Q, \vec{\vec{\gamma}}) = \prod_{i=1}^{I} P(X = x_i | \vec{\beta}_i, \vec{a}, \vec{q}_i, \vec{\gamma}_i),$$

for a vector of item responses $\vec{x} = (x_1,...,x_I)$, a Q-matrix *Q*, and a skill profile $\vec{a} = (a_1,...,a_K)$, as well as matrix-valued item difficulties $\vec{\vec{\beta}}$ and slopes $\vec{\vec{\gamma}}$. The marginal probability of a response vector is given by

$$P(\vec{X} = \vec{x} | \vec{\vec{\beta}}, Q, \vec{\vec{\gamma}}) = \sum_{\vec{a}} \pi_{\vec{a}} \prod_{i=1}^{I} P(X = x_i | \vec{\beta}_i, \vec{a}, \vec{q}_i, \vec{\gamma}_i),$$

which is the sum over all skill patterns $\vec{a} = (a_1,...,a_K)$, assuming that the discrete count density of the skill distribution is $\pi_{\vec{a}} = p(\vec{A} = \vec{a})$.

De la Torre and Douglas (2004) estimated the dichotomous version of this model, the linear logistic model (LLM; Hagenaars, 1993; Maris, 1999), using Markov chain Monte-Carlo (MCMC) methods. For ordinal skills with s_k levels, the a_k may be defined using $a(x) = x$ for $x = 0,..,(s_k - 1)$ or $a(0) = -s_k/2,..,a(s_k - 1) = s_k/2$. The parameters of the models as given in Equation 2 can be estimated for dichotomous and polytomous data, as well as for ordinal skills, using the EM algorithm.

An Example of a Simple Diagnostic Model

How do diagnostic models look? In the following example, there are two hypothesized skills, a dichotomous mastery/non-mastery skill $T1 \in \{-1,1\}$ and an ordinal skill, $T2 \in \{-2,-1,0,1,2\}$, with five proficiency levels.

In addition, there are seven observed variables, referred to as the item response variables in psychometric models and models for educational measurement. In this example, we assume that a mixed format set of three dichotomous items, $X1...3 \in \{0,1\}$, and four polytomous items, $X4...X7 \in \{0,1,2,3\}$, is observed.

The Q-matrix, which relates items to the underlying skill variables, has two columns, one each for the two skills T1 and T2, and seven rows. The Q-matrix for this example may look like

$$Q = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \tag{3}$$

which indicates that skill T1 is required for items X1, X2, X3, and X5, but not for the remaining items. Skill T2 is required for items X2, X4, X5, X6, and X7, but not for items X1 and X3. An illustration of Equation 3 is shown in Figure 1.

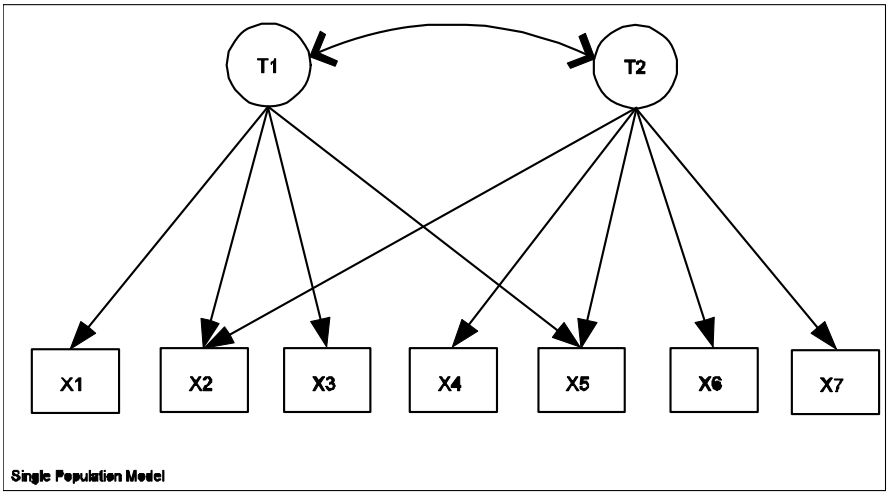

Figure 1. **A graph of the example diagnostic model.**

Models such as the one depicted in Figure 1 implicitly assume that the same structure holds for all examinees in the population from which the observations are sampled. More specifically, when using the Q-matrix given in Equation 3 and the model equation given in Equation 2, one assumes that this structure holds for all examinees in the population.

Mixtures of General Diagnostic Models

The GDM can also be extended also to a mixture-distribution IRT model (von Davier & Rost, 2006; see also the corresponding sections below). This allows for the estimation of this class of diagnostic model in different latent classes without prespecifying which observation belongs to which class and provides the ability to check whether the same kind of skill-by-item relationships holds for all the subjects sampled from a particular population. A multiple-group version of the GDM can also be specified and estimated using the algorithm described below. This allows the estimation of diagnostic models that contain partially missing grouping information (similar to the approach described in von Davier & Yamamoto, 2004b). For diagnostic models involving multiple observed groups or multiple unobserved populations (latent classes), parameter constraints can be specified that ensure scale linkages across these populations. The MGDM is

$$P(X_i = x \mid \vec{\vec{\beta}}_i, \vec{a}, \vec{q}_i, \vec{\vec{\gamma}}_i, g) = \frac{\exp\left[\beta_{xig} + \sum_{k=1}^{K} x\gamma_{ikg} q_{ik} \theta(a_k)\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yig} + \sum_{k=1}^{K} y\gamma_{ikg} q_{ik} \theta(a_k)\right]},$$

with parameters as defined above and added group index g. This model allows the estimation of separate model parameters in the *g* separate groups. The groups may be defined by an observed-group indicator variable; in this case, the above model is the diagnostic model equivalent of a multiple-group IRT model (Bock & Zimowski, 1997). If the groups are unobserved and have to be inferred during estimation, the above model is a discrete mixture diagnostic model (see von Davier & Rost, 2006; von Davier & Yamamoto, 2007). [

The marginal probability of a response vector in the MGDM is

$$P(\vec{X} = \vec{x} \,|\, \vec{\vec{\vec{\beta}}}, Q, \vec{\vec{\gamma}}) = \sum_g \pi_g \sum_{\vec{a}} \pi_{\vec{a}|g} \prod_{i=1}^{I} P(X_i = x \,|\, \vec{\vec{\beta}}_i, \vec{a}, \vec{q}_i, \vec{\vec{\gamma}}_i, g),$$

with cube-valued (classes *g* times items *i* times categories *x*) item difficulties,

$$\vec{\vec{\beta}} = (\beta_{gik})_{g=1..G;i=1..I;k=1..K},$$

and cube-valued (classes g times items i times skills k) slope parameters,

$$\vec{\vec{\gamma}} = (\gamma_{gik})_{g=1..G;i=1..I;k=1..K},$$

and a (0/1) Q-matrix Q. Let π_g denote the relative class or group sizes and $\pi_{\vec{a}|g} = P(\vec{a} \mid g)$ denote the class- or group-specific distribution of skill patterns $\vec{a}$ in group g.

Figure 2 illustrates a diagnostic model in multiple populations. This figure indicates that the item parameters and skill distributions are modelled separately in the different instances of the grouping variable g by providing a separate graph for each group.

Instead of separate graphs for separable populations, groups, or classes, the dependency of the model parameters on the group indicator can be illustrated by adding some arrows to the diagnostic model graph in Figure 1. Figure 3 presents the multiple-group variant in this manner. In this figure, the circles in shades of gray targeted by the *Group?Class* population indicator variable represent the dependency of item parameters on the population.

In Figure 3, all arrows originating from latent variables T1 and T2 include a gray circle that indicates the population dependency of the item parameters. Originating from a new variable (Group?Class) are the arrows that target these ‘population dependency’ indicators. In addition, the distribution of latent variables T1 and T2 are on the receptive end of arrows originating from the group indicator variable (Group?Class), indicating that the latent trait distributions for T1 and T2 may also vary across populations.

Multiple-group and mixture models for item response data may be used to study how different two or more populations are by looking at how parameter sets from the model differ for different groups or subpopulations. These models are useful to separate samples into groups employing different strategies to solve items (Kelderman & Macready, 1990; Rost & von Davier, 1993). Researchers have used these models to identify response styles and faking (Eid & Zickar, 2007; Rost, Carstensen, & von Davier, 1996, 1999). Rijmen and DeBoeck (2003) studied the relationship of mixture IRT models to MIRT (Reckase, 1985). More generally, mixture models can be used to test whether a unidimensional IRT model is appropriate for the data at hand (Rost & von Davier, 1995).

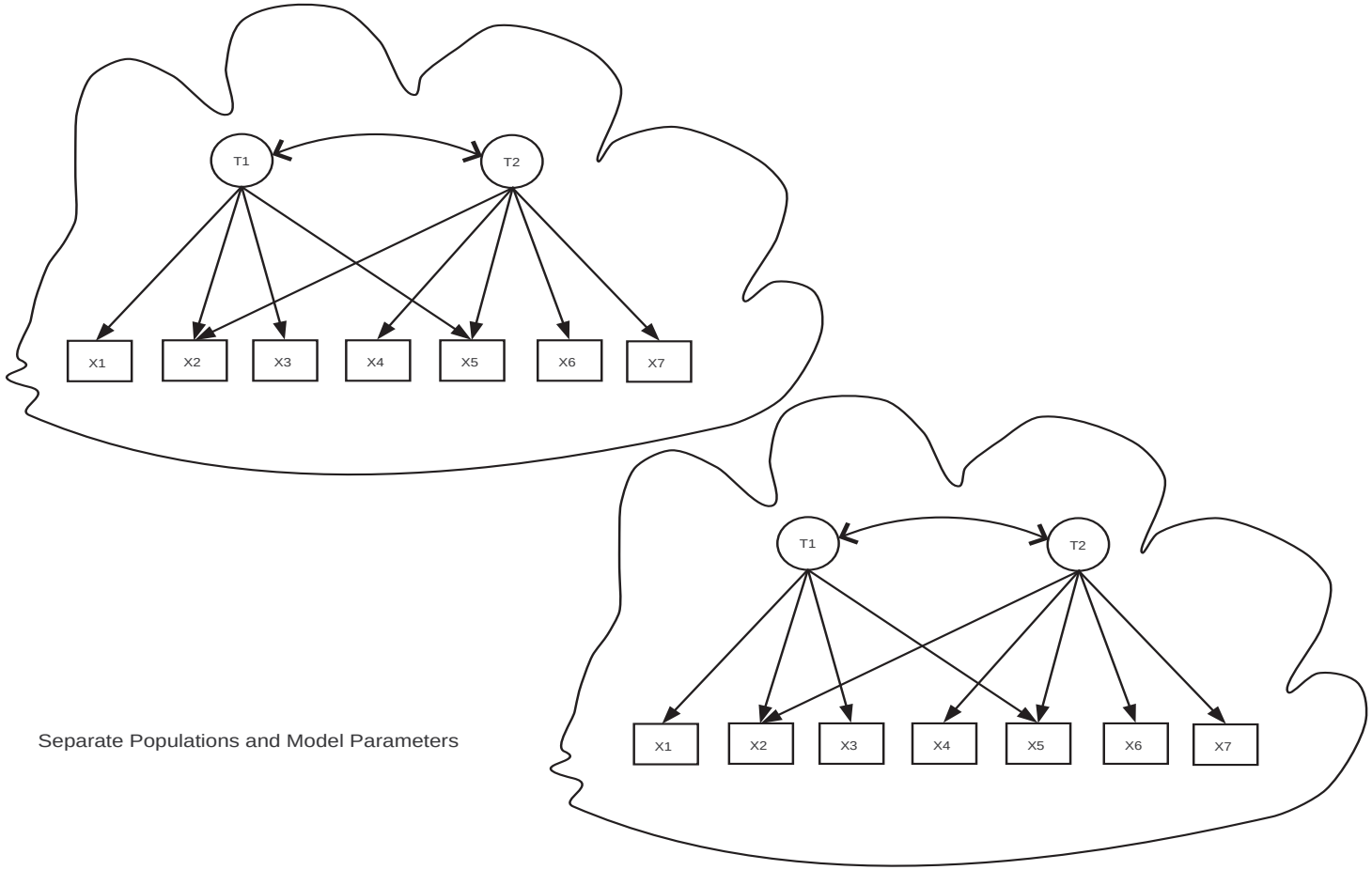

Figure 2. **A graph of a multiple-population or mixture diagnostic model.**

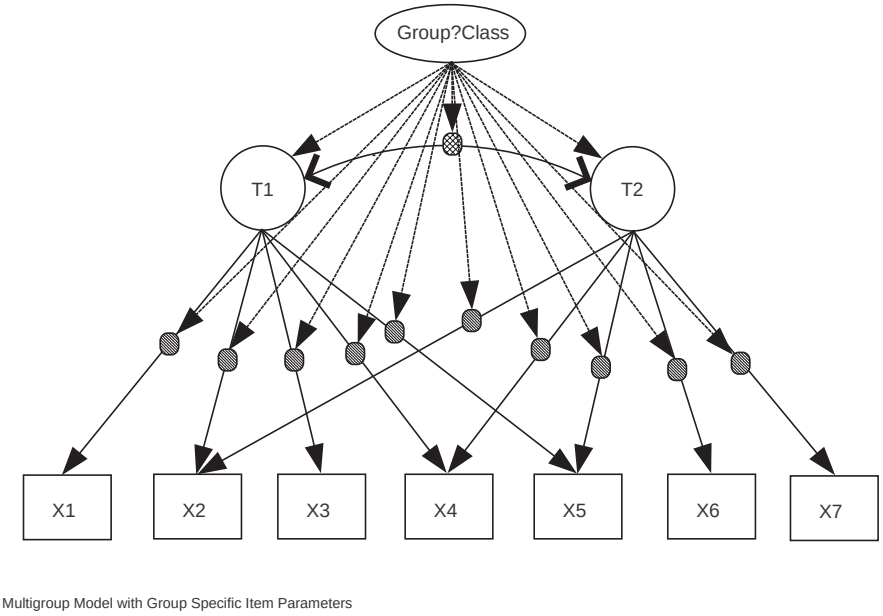

Figure 3. **An alternative graph of a multiple-group or mixture diagnostic model.**

Scale Linkages Across Mixture/Multiple-Group Diagnostic Models

In the unconstrained case, all item parameters may differ across the subpopulations in an MGDM. This is not always desirable because comparisons across groups require some common interpretation of the parameters involved. This is usually interpreted as meaning the parameters have to be on the same IRT scale. Scale linkage in IRT models enables the comparison of ability estimates across different populations (see, e.g., von Davier & von Davier, 2004; Kolen & Brennan, 1995). The scale indeterminacy of IRT models makes these models invariant under appropriate linear transformations of the parameters involved so that parameter estimates of common items can be transformed (or constrained) to match certain objectives. The objective to be met is either matching moments of the item parameters shared across forms or populations or setting equal the common items' parameters across forms or populations (von Davier & von Davier). This objective can be accomplished by employing constrained maximum likelihood estimation or by maximizing a modified likelihood that adds a penalty term or a Lagrange multiplier (Aitchison & Silvey, 1958).

In MGDMs, comparisons across subpopulations are made possible in the same way items are constrained in IRT scale linkages. The most stringent comparisons are made possible by assuming that the same item parameters hold for a set of common items across subpopulations. In a graph, arrows originating from the group indicator mean that the

targeted parameter depends on the group indicator, while the absence of arrows mean that the targeted parameter is independent of the grouping indicator g (i.e., the absence of arrows indicates that constraints on item parameters are to be equal across subpopulations).

Figure 4 illustrates this sort of linkage in MGDMs. The items without a direct arrow originating from the ellipse labelled *group* are items X2, X4, and X5; these items have the same parameters across subpopulations. Items X1, X3, X6, and X7 have group-dependent parameters, which are indicated by arrows originating from the *group* ellipse. The same holds for the skill variables $\theta 1$ and $\theta 2$ as well as their covariance. The distribution of these variables and their relationship do not vary across subpopulations, which is indicated by an arrow originating from the *group* ellipse.

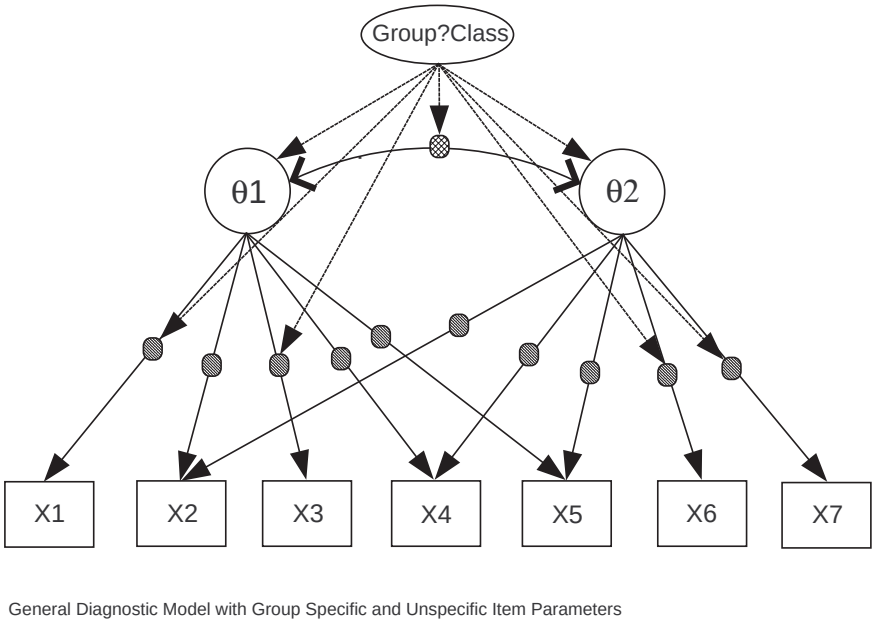

Figure 4. **A mixture/multiple-group general diagnostic model with equality constraints.**

The *mdltm* software (von Davier, 2005b) allows for the definition of equality constraints across pairs of items or multiple items in different subpopulations, as well as constraints that affect only difficulties or slopes in MGDMs. In addition, parameter constraints can be employed that fix item parameters to certain values, for example, to parameter values from previous calibrations. Other scale linkages such as the *mean-variance* methods used in unidimensional IRT (Loyd & Hoover, 1980; Marco, 1977) are also available for estimating linked GDMs in several populations by invoking corresponding key words in the *mdltm* scripting language.

In addition to scale-linkage methods that mirror traditional methods used in IRT, parameter constraints for MGDMs can be used to develop new methods of scale linkage and even new models within the class of MGDMs as outlined next.

Different Q-Matrices in Different Populations

The same skill-by-item structure implied by the Q-matrix might not be appropriate for all subpopulations. Imagine that different student groups receive test preparation from different vendors, so that some students are trained to use additional methods to make sure their responses are correct. In this case, different Q-matrices might hold in different subpopulations since some student groups are trained to use this additional method, and may do so to with varying success. The other students most likely do not know about this method, since they have not been trained in using it.

In the framework of MGMs, this can be implemented as follows using the methods of parameter constraints offered by *mdltm:* Define a super Q-matrix with entries of 1 if a skill is needed for an item in at least one subpopulation and set the Q-matrix to 0 only if the skill is not required for an item in all subpopulations. Then fix slope parameters to equal 0 for skills that are not needed in certain subpopulations for certain items. This ensures that no slope is estimated in these subpopulations as the slope has been set to equal 0. In these subpopulations, the corresponding skill (with the slope equalling 0 for certain items) does not contribute to items constrained in that way.

In the next step, the fit of these constrained models with unique Q-matrices across subpopulations may be compared to models that do not impose such constraints. This will provide evidence on how appropriate are the assumptions that lead to a specific constrained model.

Strongest Form of Linkage Across Multiple Populations

Another important case of a constrained model for multiple populations is a multiple-group model where all (common) items are assumed to have the same parameters in all subpopulations. This means that while each common item may have a parameter that differs from other items in the same population, the common item is assumed to have the same parameters across populations (i.e., different administrations, different cohorts). Only the ability distributions differ across subpopulations. For instance, the ability distribution in the example is $P(\theta 1, \theta 2 \mid g)$, where g stands for the group or population under consideration.

This form of the model measures identical skills, allowing for different skill distributions across subpopulations.

The rationale for a multiple-group model that includes just one set of item parameters is the assumption of measurement invariance, in the sense that the item's conditional response probability depends on a unidimensional, person-specific variable only. Given the value of this variable (e.g., the skill, ability, or proficiency of an examinee) and knowing the item characteristics or parameters, the response probability is determined without respect to which group the examinee belongs. Figure 5 illustrates this form of equality constraint across subpopulations; note that the arrows originating from the *group* ellipse target the skill distribution variables only, and no arrow targets the gray bubbles, which represent the item characteristics.

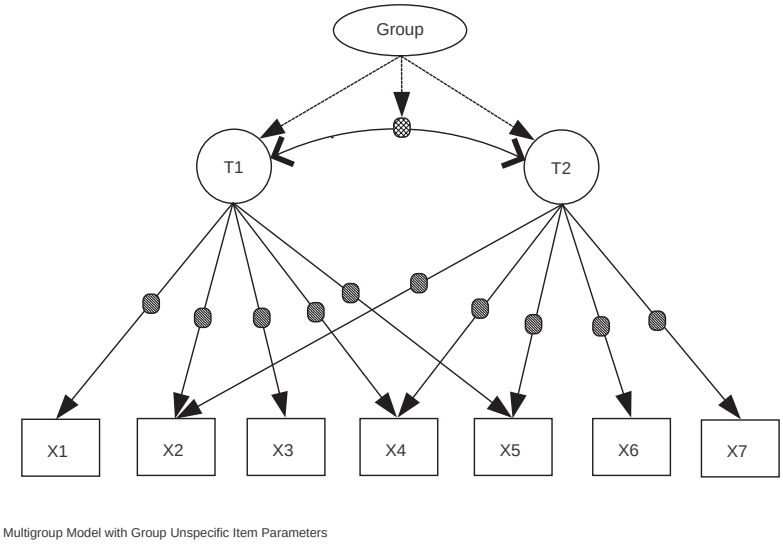

Figure 5. **Strongest form of linkage across multiple populations.**

This measurement-invariant MGDM assumes that item characteristics are exactly the same across groups, while allowing skill distribution differences across groups. Other models are easily obtained by varying the types of constraints presented in this paper.

Applications

GDMs have been applied in research studies conducted for different large scale assessment programs. This section gives a brief summary of three applications to data from such programs.

General Diagnostic Models and English Language Testing

von Davier (2005a) analyzed TOEFL® iBT pilot data with common IRT models and GDMs using expert-generated Q-matrices. It was hypothesized that four skills each could be identified in the listening and in the reading sections of the TOEFL iBT pilot data. In contrast to this expectation, it was found that a unidimensional mixed two-parameter logistic (2PL)/GPCM IRT model fits the data as well as the GDM with four dichotomous mastery/nonmastery skills.

For reasons of parsimony, von Davier (2005a) concluded that the 2PL/GPCM was to be favored for both the listening and the reading section. Figure 6 shows the relation between the 2PL ability estimate and the skill-mastery probabilities for the listening section data of this study. Figure 6 shows very similar results for the two forms of the TOEFL iBT, Form A and Form B, used in von Davier's study. It is evident that all four skills have a rather strong relationship to the overall 2PL parameter estimate. The probability of skill mastery increases in a very systematic fashion with increasing 2PL parameter. The width of the four S-shaped plots is mainly a function of reliability of the skill-mastery probability. If the skill is measured by many items, the S-shaped curve is narrower; if few items are used to measure the skill, the S-shape is a little wider.

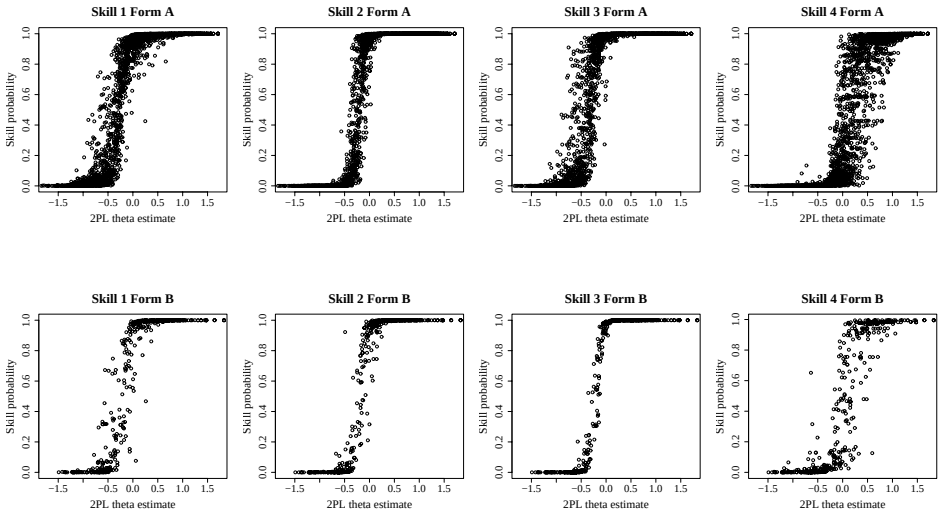

Figure 6. **English language general diagnostic model, listening Forms A and B.**

In additional analyses, *mdltm* was used to test a unidimensional IRT model, a two-dimensional IRT model employing the 2PL/GPCM, and a GDM. Each model contained all eight skills (four for reading, four for listening) in one Q-matrix and was composed of the joint listening and reading parts of the TOEFL iBT pilot data. It was found that the two-

dimensional discrete IRT model estimated in the GDM framework provides the best data description in terms of balancing parsimony and model-data fit.

MGDMs for Matrix Samples of Item Responses

Xu and von Davier (2006) used a multiple-group GDM with large-scale survey data. In their example, gender and race/ethnicity were used as grouping variables. Xu and von Davier used data from the 2002 12th-grade National Assessment of Educational Progress (NAEP; for the history of the national assessment, see Jones & Olkin, 2004) in reading and mathematics. The reading data were modeled using MGDMs with up to three dimensions, and the mathematics data were modeled using MGDMs with up to seven dimensions (four content domains plus three complexity levels). because data from large-scale surveys are extremely sparse in nature, the authors performed a parameter-recovery study based on estimating GDMs in sparse samples of item responses.

The results are reported in detail in Xu and von Davier (2006). The parameter-recovery results under different levels of sparseness of data support the feasibility of estimating GDMs under such conditions. Table 1 presents results of this study, making use of the average bias and the root mean square error obtained under different degrees of data sparseness.

Table 1

Bias and RMSE of GDM Item Difficulties and Slope Parameter and Skill Distribution Probability Estimates

		Percentage of missing data		
	Measure	10%	25%	50%
Item parameters	Average bias	0.001	0.002	0.005
	Average RMSE	0.071	0.083	0.119
Skill distribution	Average bias	0.000	0.000	0.000
	Average RMSE	0.004	0.004	0.007

The results reported by Xu and von Davier (2006) on the NAEP data showed that a multidimensional MGDM (both single-group and multiple-group versions of the GDM were tested) was found to fit the reading data consistently better than a unidimensional IRT model. However, a unidimensional IRT model fit the math data better than a three-, four-, or even a seven-dimensional GDM. This result has since been replicated using other larger NAEP data

sets and can be explained by the fact that the reading domains correlate less than do the math subscales when defined by either content or complexity factors.

Mixture IRT, General Diagnostic, and Latent Class Models

Huang and von Davier (2006) used background data from an international large-scale assessment of adult literacy. Their results were based on data from approximately 47,000 adults assessed with cognitive adult literacy scales and background questionnaires. The sample contained data from seven countries.

The goal of this study was to develop indicator variables using latent class models, GDMs, or common IRT models. The purpose of these derived indicator variables was to provide more reliable background variables for secondary data analysis by aggregating response data.

In this study, the three models listed above fit the relatively short scales equally. Measures of model-data fit made it evident that a distinction between discrete and continuous latent variables was difficult to make if only a few observed variables were used.

Conclusions and Outlook

This paper introduces MGDMs and presents evidence for the utility of this class of models. So far, examples of successful applications of the GDM and its mixture generalizations come from data analyses aimed at identifying the necessary level of complexity needed to fit observed responses and exploring multiple-group versus single-group models as examples of scale linkages across multiple populations.

Obvious next steps include the introduction of covariates for predicting skill distributions. One common way to do this is to extend the GDM using a latent regression model—a conditioning model in the language of NAEP and other large-scale survey assessments (von Davier, Sinharay, Oranje, & Beaton, 2006). Figure 7 illustrates this model extension with the example used in previous figures.

Xu and von Davier (2006) developed parametric skill-distribution models for GDMs. These parametric families of discrete skill distributions enable the skill space to be modeled more parsimoniously, so that models with a larger skill count are still estimable even when the sample sizes in the different subpopulations are not large. These extensions have been implemented in *mdltm* and can be estimated with customary maximum likelihood methods. These developments are currently being studied using a variety of large-scale data sets.

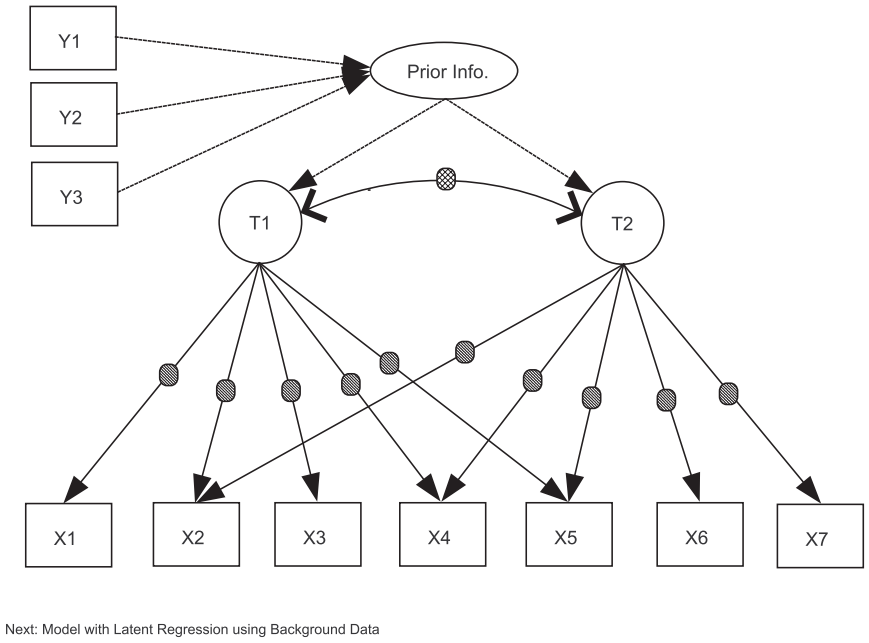

Figure 7. **Extending the GDM using a latent regression model.**

However, even with computationally efficient methods and fast computers that enable complex models to be estimated in a reasonable time, the question of model complexity remains important. For this reason, research on model-data fit and the balance between parsimony of models is imperative (see Haberman & von Davier, 2006).

Mixture general diagnostic models are a useful tool for educational measurement research. The potential of these models for practical large-scale data analysis lies in the fact that models of different complexity can be specified within a common framework, estimated using standard maximum likelihood methods, and directly compared in terms of their predictive power.

References

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813–829.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag.

De la Torre, J., & Douglas, J. A. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), C*ognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

Eid, M., & Zickar. M. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255–270). New York: Springer-Verlag.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87–111.

Goodman. L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231.

Goodman, L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology, 79,* 1179–1259.

Haberman, S. J. (1979). Qualitative data analysis: Vol. 2. New developments. New York: Academic Press.

Haberman, S. J., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Hagenaars, J. A. (1993). Quantitative applications in the social sciences: Vol. 94. Loglinear models with latent variables. Newbury Park, CA: Sage.

Hartz, S., Roussos, L., & Stout, W. F. (2002). *Skills diagnosis: Theory and practice* [Computer software manual for Arpeggio]. Princeton, NJ: ETS.

Huang, X., & von Davier, M. (2006, April). *Comparing latent trait models for large-scale survey background data.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307–327.

Kolen, M. J., & Brennan, R. J. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120.

Marco, G. L. (1977). Item characteristic curves solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187–212.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Rijmen, F., & De Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research, 67*(3), 219–231.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Rost, J., Carstensen, C. H., & von Davier, M. (1996). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences.* Münster, Germany: Waxmann.

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-Skalierbar? *Diagnostica, 45*(3), 119–127.

Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology. Proceedings of the 7th European meeting of the Psychometric Society in Trier.* Stuttgart, Germany: Gustav Fischer Verlag.

Rost, J., & von Davier, M. (1995). *Mixture distribution Rasch models.* In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models—Foundations, recent developments and applications* (pp. 257–268). New York: Springer.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2005b). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.

von Davier, M., DiBello, L., & Yamamoto, K. Y. (2006). *Reporting test outcomes with models for cognitive diagnosis* (ETS Research Rep. No. RR-06-28). Princeton, NJ: ETS.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York: Springer-Verlag.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26: Psychometrics*. Amsterdam: Elsevier.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Marginal estimation of population characteristics: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics.* Amsterdam: Elsevier.

von Davier, M., & von Davier, A. A. (2004). *A unified approach to IRT scale linkage and scale transformations* (ETS Research Rep. No. RR-04-09). Princeton, NJ: ETS.

von Davier, M., & Yamamoto, K. (2004a, October). *A class of models for cognitive diagnosis.* Paper presented at the fourth Spearman conference, Philadelphia.

von Davier, M., & Yamamoto, K. (2004b). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389–406.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution models and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–118). New York: Springer-Verlag.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-06-08). Princeton, NJ: ETS.